



(51) International Patent Classification:

H04L 12/729 (2013.01) H04L 12/753 (2013.01)
H04L 12/707 (2013.01)

(21) International Application Number:

PCT/US2012/065115

(22) International Filing Date:

14 November 2012 (14.11.2012)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

61/560,226 15 November 2011 (15.11.2011) US
13/648,961 10 October 2012 (10.10.2012) US

(71) Applicant: ORACLE INTERNATIONAL CORPORATION [US/US]; 500 Oracle Parkway, M/S 50P7, Redwood Shores, California 94065 (US).

(72) Inventors: GUAY, Wei, Lin; 25-12-1, Parkview Tower, Tingkat Bukit Jambul 1, Bayan Lepas, Penang, 11950 (MY). JOHNSEN, Bjorn-Dag; Vilberggrenda 9, N-0687 Oslo (NO).

(74) Agents: MEYER, Sheldon, R. et al.; FLIESLER MEYER LLP, 650 California Street, Fourteenth Floor, San Francisco, California 94108 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM,

AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

- with international search report (Art. 21(3))
- before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))

(54) Title: SYSTEM AND METHOD FOR USING DYNAMIC ALLOCATION OF VIRTUAL LANES TO ALLEVIATE CONGESTION IN A FAT-TREE TOPOLOGY

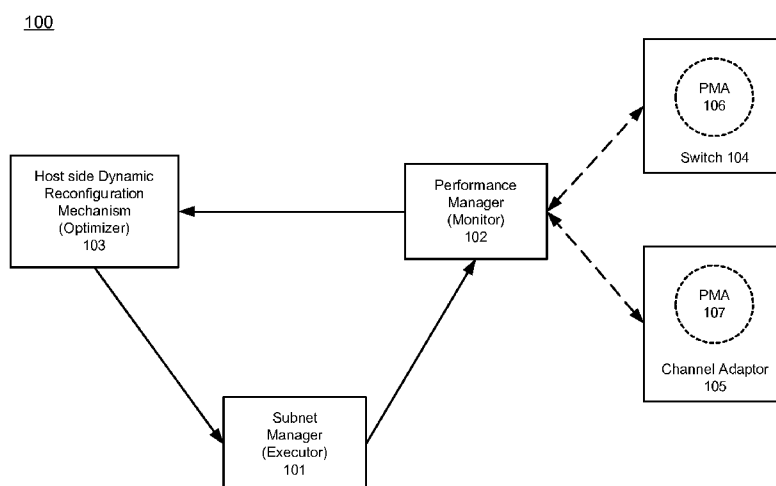


FIGURE 1

(57) Abstract: A system and method can prevent traffic congestion in a middleware machine environment with a plurality of switches in a fat-tree topology. A subnet manager can sweep a subnet in the middleware machine environment to discover changes and maintain the subnet fully connected. A performance manager can retrieve performance and error-related information from one or more performance management agents that are associated with one or more components in the subnet. Then, a host can dynamically reconfigure one or more virtual lanes in order to improve network performances.

SYSTEM AND METHOD FOR USING DYNAMIC ALLOCATION OF VIRTUAL LANES TO ALLEVIATE CONGESTION IN A FAT-TREE TOPOLOGY

Copyright Notice:

5 **[0001]** A portion of the disclosure of this patent document contains material which is subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by anyone of the patent document or the patent disclosure, as it appears in the Patent and Trademark Office patent file or records, but otherwise reserves all copyright rights whatsoever.

10 **Field of Invention:**

[0002] The present invention is generally related to computer systems, and is particularly related to preventing head-of-line blocking and traffic congestion in a middleware machine environment.

15 **Background:**

[0003] The interconnection network plays a beneficial role in the next generation of super computers, clusters, and data centers. High performance network technology, such as the InfiniBand (IB) technology, is replacing proprietary or low-performance solutions in the high performance computing domain, where high bandwidth and low latency are the key requirements. For example, IB installations are used in supercomputers such as Los Alamos National Laboratory's Roadrunner, Texas Advanced Computing Center's Ranger, and Forschungszentrum Juelich's JuRoPa.

20 **[0004]** IB was first standardized in October 2000 as a merge of two older technologies called Future I/O and Next Generation I/O. Due to its low latency, high bandwidth, and efficient utilization of host-side processing resources, it has been gaining acceptance within the High Performance Computing (HPC) community as a solution to build large and scalable computer clusters. The de facto system software for IB is OpenFabrics Enterprise Distribution (OFED), which is developed by dedicated professionals and maintained by the OpenFabrics Alliance. OFED is open source and is available for both GNU/Linux and Microsoft Windows.

30

Summary:

[0005] Described herein is a system and method that can prevent head-of-line blocking and traffic congestion in a middleware machine environment with a plurality of switches in a fat-tree topology. A subnet manager can sweep a subnet in the middleware machine environment to discover changes and maintain the subnet fully connected. A performance manager can retrieve performance and error-related information from one or more performance management agents that are associated with one or more components in the subnet. Then, a host can dynamically

35

reconfigure one or more virtual lanes in order to improve network performances.

Brief Description of the Figures:

5 [0006] **Figure 1** shows an illustration of optimization feedback cycle in a middleware environment in accordance with an embodiment of the invention.

[0007] **Figure 2** illustrates an exemplary flow chart for alleviating network congestion in a middleware environment in accordance with an embodiment of the invention.

[0008] **Figure 3** shows an illustration of dynamic allocation of virtual lanes to alleviate congestion in a fat-tree topology in accordance with an embodiment of the invention.

10 [0009] **Figure 4** shows an illustration of dynamic allocation of virtual lanes to alleviate congestion in an over-subscribed fat-tree topology in accordance with an embodiment of the invention.

Detailed Description:

15 [0010] Algorithmic predictability of network traffic patterns is reduced with the introduction of virtualization and many-cores systems. When multiple virtualized clients reside on the same physical hardware, the network traffic becomes an overlay of multiple traffic patterns that might lead to hot-spots in the network. A hot-spot occurs if multiple flows are directed toward a single endpoint. Common sources for hot-spots include complex traffic patterns due to virtualization,
20 migration of virtual machine images, checkpoint and restore mechanisms for fault tolerance, and storage and I/O traffic.

[0011] When a hot-spot exists in a network, the flows designated for the hot-spot might reduce the performance for other flows, called victim flows, not designated to the hot-spot. This is due to the head-of-line (HOL) blocking phenomena created by the congested hot-spot.

25 [0012] One way to avoid this problem is to use a congestion control (CC) mechanism such as the CC mechanism evaluated in hardware. However, the congestion control mechanism evaluated in hardware may not always be available, e.g. due to a mixture of old and new equipments coexisting in large clusters. Furthermore, the selection of the appropriate CC parameters highly depends on the topology and incorrect parameters might lead to performance
30 degradation. Additionally, some oscillations can occur among the flows due to the fact that the congestion control mechanism is dynamically adjusting the injection rate of the senders.

[0013] In accordance with an embodiment of the invention, a system and method can prevent head-of-line blocking and traffic congestion in an interconnected network, such as a middleware machine environment with a plurality of switches using a fat-tree topology. A subnet manager
35 can sweep a subnet in the middleware machine environment to discover changes and maintain the subnet fully connected. A performance manager can retrieve performance and error-related

information from one or more performance management agents that are associated with one or more components in the subnet. Then, a host can dynamically reconfigure one or more virtual lanes in order to improve network performances.

5 InfiniBand (IB) Architecture

[0014] In accordance with an embodiment of the invention, traffic congestion can be prevented in the InfiniBand (IB) architecture, which is a serial point-to-point technology. Each of the IB networks, or subnets, can include a set of hosts interconnected using switches and point-to-point links. A single subnet is scalable to more than ten-thousand nodes and two or more subnets can be interconnected using an IB router. The hosts and switches within a subnet are addressed using local identifiers (LIDs), e.g. a single subnet is limited to 48151 unicast addresses.

[0015] An IB subnet can employ at least one subnet manager (SM) which is responsible for initializing and starting up the sub-net including the configuration of all the IB ports residing on switches, routers and host channel adapters (HCAs) in the subnet. The SM's responsibility also includes routing table calculation and deployment. Routing of the network aims at obtaining full connectivity, deadlock freedom, and load balancing between all source and destination pairs. Routing tables can be calculated at network initialization time and this process can be repeated whenever the topology changes in order to update the routing tables and ensure optimal performance.

[0016] At the time of initialization, the SM starts in the discovering phase where the SM does a sweep of the network in order to discover all switches and hosts. During the discovering phase, the SM may also discover any other SMs present and negotiate who should be the master SM. When the discovering phase is completed, the SM can enter a master phase. In the master phase, the SM proceeds with LID assignment, switch configuration, routing table calculations and deployment, and port configuration. At this point, the subnet is up and ready to use.

[0017] After the subnet is configured, the SM can monitor the network for changes (e.g. a link goes down, a device is added, or a link is removed). If a change is detected during the monitoring process, a message (e.g. a trap) can be forwarded to the SM and the SM can reconfigure the network. Part of the reconfiguration process, or a heavy sweep process, is the rerouting of the network which can be performed in order to guarantee full connectivity, deadlock freedom, and proper load balancing between all source and destination pairs.

[0018] The HCAs in an IB network can communicate with each other using Queue Pairs (QPs). A QP is created during the communication setup, and a set of initial attributes such as QP number, HCA port, destination LID, queue sizes, and transport service are supplied. On the

other hand, the QP associated with the HCAs in a communication is destroyed when the communication is over. An HCA can handle many QPs, each QP consists of a pair of queues, a Send Queue (SQ) and a Receive Queue (RQ). There is one such pair present at each end-node that is participating in the communication. The send queue holds work requests to be transferred
5 to the remote node, while the receive queue holds information on what to do with the data received from the remote node. In addition to the QPs, each HCA can have one or more Completion Queues (CQs) that are associated with a set of send and receive queues. The CQ holds completion notifications for the work requests posted to the send and receive queue.

[0019] The Subnet Administrator (SA) is a subnet database associated with the master SM to
10 store different information about a subnet. The communication with the SA can help the end-node to establish a QP by sending a general service management datagram (MAD) through a designated QP, e.g. QP1. Both sender and receiver require information such as source/destination LIDs, service level (SL), MTU, etc. to establish a QP. This information can be retrieved from a data structure known as a path record that is provided by the SA. In order to
15 obtain a path record, the end-node can perform a path record query to the SA, e.g. using the *SubnAdmGet/SubnAdmGetable* operation. Then, the SA can return the requested path records to the end-node.

[0020] The SM is also responsible for monitoring the network for changes using Subnet Management Agents (SMAs) that are presented in every switch and/or every HCA. The SMAs
20 communicate changes, such as new connections, disconnections, and port state change, to the SM using traps and notices.

[0021] A trap is a message sent to alert end-nodes about a certain event. A trap can contain a notice attribute with the details describing the event. Different traps can be defined for different events. In order to reduce the unnecessary distribution of traps, IB applies an event forwarding
25 mechanism where end-nodes are required to explicitly subscribe to the traps they want to be informed about.

An Optimization Feedback Cycle for Performance Management

[0022] **Figure 1** shows an illustration of optimization feedback cycle in a middleware
30 environment in accordance with an embodiment of the invention. As shown in Figure 1, an optimization feedback cycle in the middleware environment includes an executor (e.g. a subnet manager 101), a monitor (e.g. a performance manager 102), and an optimizer (e.g. a switch 104).

[0023] The subnet manager 101 can periodically sweep a subnet to discover changes and to
35 maintain a fully connected subnet. Furthermore, the performance manager 102 can periodically collect information from every component in the subnet in order to analyze the network

performance, and the host side stack 103 can dynamically reconfigure the addressing state information for network configurations.

[0024] Additionally, each device in the subnet, such as a switch 104 or a channel adapter 105, can implement a performance management agent (PMA) 106 or 107. Each PMA can be associated with a set of performance monitoring and error monitoring registers. The performance manager 102 can retrieve performance and error-related information from these registers, for example using a performance management datagram (MAD).

[0025] Performance management is one of the general management services provided by IB to retrieve performance statistics and error information from IB components. Each IB device can implement a PMA and a minimum set of performance monitoring and error monitoring registers. In addition, the IB specification also defines a set of optional attributes permitting the monitoring of additional performance and error counters.

[0026] The performance manager (PM) can retrieve performance and error-related information from these registers, by issuing a performance MAD to the PMA of a given device. The PM then executes the retrieval and returns the result to the PMAs. The PM can use this information to detect incipient failures and based on this information, the PM can advise the SM about recommended or required path changes and performance optimizations.

[0027] Performance management is related to performance tuning, including finding and eliminating bottlenecks. The optimization feedback cycle as shown in Figure 1 can be applied to support using dynamic allocation of virtual lanes to alleviate network congestion with the help of the SM, the PM, and host stack with the host side dynamic reconfiguration capability. In a subnet, the SM periodically sweeps the subnet to discover changes and maintain a fully connected subnet. The PM can periodically collect information from every component in the subnet in order to analyze the network performance. After the analysis, the PM forwards the relevant information to host stack that reconfigures the virtual lanes in order to improve network performance.

[0028] In accordance with an embodiment of the invention, a routing algorithm can utilize multiple virtual lanes (VLs) to improve performance during the existence of hot-spots. The VLs can be assigned statically during the routing table generation and can avoid the negative impact of the congestion, with the assumption that the topology is a balanced, fully populated and fault-free fat-tree. Additionally, a mechanism using dynamic allocation of virtual lanes to alleviate network congestion can be designed to identify the hot-spot flows and assign the virtual lanes dynamically.

[0029] Compared to IB congestion control, using dynamic allocation of virtual lanes to alleviate network congestion, the need for source throttling of the contributors is removed. Furthermore, the IB CC parameters can cause oscillations among all the flows, because IB CC

can dynamically adjust the injection rate of the senders. As a result, the IB CC solution might not be suitable for congestion problem of a more persistent nature because the oscillations can reduce the overall network throughput. Such persistent congestion problems occur when traffic has been moved away from a failed link, when multiple jobs run on the same system and
5 compete for network resources, or when a system is not balanced for the application that runs on it. The persistent congestion problems can be handled by first detecting them, and thereafter dynamically redistributing the VL resources so as to obtain a balance that may be impossible to achieve statically at system start-up.

[0030] In accordance with an embodiment of the invention, a SM can be used along with the
10 PM enabled. The added overhead due to that the PM periodically queries the performance counters in each component within the subset can have minimal impact on data traffic, as long as the SM is running on a dedicated node.

[0031] In accordance with an embodiment of the invention, the subnet in the middleware machine environment is in a fat-tree topology. Preferably, the fat-tree topology in the middleware
15 machine environment is a simple fat-tree topology. Preferably, the fat-tree topology in the middleware machine environment is an over-subscribed fat-tree topology.

[0032] In accordance with an embodiment of the invention, one or more switches in the middleware machine environment use a routing algorithm to ensure deadlock freedom for a packet forwarded from a source node to a destination node.

[0033] Preferably, one or more switches in the middleware machine environment allows the
20 routing algorithm to forward the packet through a least common ancestor node of the source node and the destination node. Preferably, the host allows the routing algorithm to include an upward stage, in which the packet is forwarded from the source node to the least common ancestor node, and a downward stage, in which the packet is forwarded from the least common
25 ancestor node to the destination.

[0034] Preferably, one or more switches in the middleware machine environment configures one or more physical links in the middleware machine environment to be associated with multiple virtual lanes.

[0035] Preferably, the performance manager uses one or more performance counter as
30 metrics to identify hot-spots and contributors to the hot-spots during network operation. Preferably, the performance manager may identify a hot-spot flow after a performance sweep. The performance manager may forward a repath message to all contributors after discovering a hot-spot.

[0036] The middleware machine may reassign the hot-spot flow to a virtual lane classified as
35 a slow lane. The middleware machine may allow the hot-spot flow in a simple fat-tree topology to be assigned to a virtual lane classified as a slow lane if a victim flow shares an upward stage

with the hot-spot flow. The middleware machine may allow the hot-spot flow in an over-subscribed fat-tree topology to be assigned to a virtual lane classified as a slow lane if a victim flow shares at least one of an upward stage and a downward stage with the hot-spot flow. The middleware machine may reassign the hot-spot flow to a virtual lane classified as a slow lane.

- 5 The middleware machine may direct all flow to a node from a slow lane to a fast lane after determining that a node is no longer a hot spot. The middleware machine may allow a victim flow to avoid head-of-line blocking using a congestion control mechanism.

[0037] **Figure 2** illustrates an exemplary flow chart for alleviating network congestion in a middleware environment in accordance with an embodiment of the invention. Processes shown
10 in this flow chart may be executed by a processor. As shown in Figure 2, at step 201, a subnet manager sweeps a subnet in the middleware machine environment to discover changes and maintain the subnet fully connected. Then, at step 202, a performance manager can retrieve performance and error-related information from one or more performance management agents that are associated with one or more components, which may be hardware, in the sunet. Finally,
15 at step 203, the system allows a host to dynamically reconfigure network connection, such as the addressing state information, in order to improve network performances.

Alleviate Congestion in a Fat-tree Topology

[0038] In accordance with an embodiment of the invention, the optimization feedback cycle
20 mechanism can be applied to any topology and routing algorithm. In one example, fat-trees can be used because of the simplicity with respect to freedom from deadlock. The system can dynamically updates congested connections and move congested traffic flows to a different virtual lane in the fabric, and thereby ensure that the congestion effects do not have impact on connections that are not subject to the congestion.

25 **[0039]** **Figure 3** shows an illustration of dynamic allocation of virtual lanes to alleviate congestion in a fat-tree topology in accordance with an embodiment of the invention. As shown in Figure 3, the middleware machine environment 300 includes a plurality of leaf switches, e.g. switches 301 to 303, and a plurality of nodes, such as server nodes 1-6 that connect to the leaf switches in a fat-tree topology. Additionally, the leaf switches 301-303 can connect to an
30 intermediate switch or a root switch 310 using one or more physical links I-VI.

[0040] In accordance with an embodiment of the invention, each physical link can support one or more virtual lanes (VLs). The VLs are logical channels on the same physical link with separate buffering, flow control, and congestion management resources. As shown in Figure 3, each physical link I-VI can have multiple virtual lanes, such as a slow lane (virtual lane 1) and a fast
35 lane (virtual lane 0), and all packet flows can be configured to run on the fast lane initially.

[0041] A routing algorithm can ensure deadlock freedom in the fat tree topology. The routing

algorithm can include two stages: an upward stage in which the packet is forwarded from the source, and a downward stage when the packet is forward toward the destination. The transition between these two stages occurs at the least common ancestor, which is the intermediate switch or a root switch 310 that can reach both the source and the destination through its downward
5 ports.

[0042] When multiple virtualized clients reside on the same physical hardware, the network traffic becomes an overlay of multiple traffic patterns that might lead to hot-spots in the network. In the example as shown in Figure 3, end node 5 can become a hot spot, when multiple flows (in dot lines) from the contributors such as node 1, node 3, and node 6 are destined toward it.

[0043] The flows designated for a hot spot can reduce the performance for other flows. In the above example, there can be another flow from node 2 to node 3. The upward stage of the flow from node 1 to node 5 and the flow from node 2 to node 3 shares the physical link I, since physical link I is designated to handle the traffic from the leaf switch 301 to both the leaf switches 302 and 303. Due to the head-of-line (HOL) blocking phenomena, the flow from node 2 to node
15 3 can become a victim flow (in dash line).

[0044] The system can distribute the two upward stages of the two flows over different virtual lanes on the same physical link, for example by separating network flows into slow lane and fast lane traffics. After discovering that node 5 is a hot-spot, the system can trigger the forwarding of a message, e.g. a re-path trap, to all potential contributors, such as nodes 1, 3, and 6. Then, the
20 system can direct the flow from node 1 to node 5 to go through virtual lane 0 on physical link I, which is designated as the slow lane. Additionally, if a new flow is directed to the existing hot spot, node 5, the new flow can be moved to the slow lane. On the opposite, if the node 5 is no longer a hot spot, all flows directed to the node 5 can be moved back to virtual lane 1, which is classified as the fast lane on physical link I.

[0045] **Figure 4** shows an illustration of dynamic allocation of virtual lanes to alleviate congestion in an over-subscribed fat-tree topology in accordance with an embodiment of the invention. As shown in Figure 4, the middleware machine environment 400 includes a plurality of leaf switches, such as switches 401 to 403, and a plurality of nodes, such as server nodes 1-12 that connect to the leaf switches in a fat-tree topology. Additionally, the leaf switches 401-403
25 can connect to an intermediate switch or a root switch 410 using one or more physical links I-IV.

[0046] In this oversubscribed fat-tree, the downward path for forwarding a packet is shared by several destinations, instead of dedicating to a single destination as shown in Figure 3. The oversubscribed fat-tree in Figure 4 is a 2:1 oversubscribed fat-tree, since each downward path is shared by two destinations.

[0047] As shown in Figure 4, the traffic flows from nodes 1, 5, and 10 to node 9 can cause the negative impact of HOL blocking in the oversubscribed fat-tree. Thus, the hot-spot is at node 9,
35

and node 1, 5 and 10 are the contributors.

[0048] There can be two situations where the victim flows can suffer from HOL blocking when the links are oversubscribed, one at the upward stage and one at the downward stage.

[0049] As shown in Figure 4, the victim flow from node 2 to node 7 shares an upward stage from leaf switch 401 to the intermediate / root switch 410 with the hot-spot flow from the contributor node 1 to node 9, through physical link I. This is similar to the example as shown in Figure 3, where the performance reduction is due to the upstream link being shared with the congestion contributor, node 1.

[0050] Also as shown in Figure 4, the victim flow from node 2 to node 11 shares the upward link from leaf switch 401 to the intermediate / root switch 410 with the hot-spot flow from the contributor node 1 to node 9, through physical link I. Additionally, the victim flow from node 2 to node 11 shares a downward stage from the intermediate / root switch 410 to the leaf switch 403 with all hot-spot contributors. In this case, the performance reduction happens at the downstream link being shared with the congestion contributor, node 1, even though the destination node of the victim flow, node 11, is a different node from the hotspot.

[0051] The system can distribute the flows over different virtual lanes on the same physical link, for example by separating network flows into slow lane and fast lane traffics.

dFtree Algorithm

[0052] In accordance with an embodiment of the invention, a routing algorithm, e.g. the dFtree algorithm, can be use to perform the allocation of VLs dynamically during network operation using the optimization feedback cycle. A performance manager monitors the network using hardware port counters to detect congestion and optimizes the current VL allocation by classifying flows as either slow lane (contributors to congestion) or fast lane (victims of congestion). Then, the optimization can be applied using a host side dynamic reconfiguration method. The effect of this method is that all flows contributing to congestion are migrated to a separate VL (slow lane) in order to avoid the negative impact of head-of-line blocking on the flows not contributing to congestion (victim flows).

[0053] The routing algorithm can use various metrics to identify the hot-spot flows dynamically, such as IB performance counters: *XmitWait*, and *XmitData*. The IB counter *XmitWait* is the number of ticks when a port selected has data to transmit but no data was actually sent during an entire tick, e.g. because of insufficient credits or because of lack of arbitration. Here, a tick is the IB architecture hardware sampling clock interval. Furthermore, the IB counter *XmitData* is the total number of data in double words transmitted on all VLs. Additionally, the routing algorithm can use other metrics such as *Interval*, which is the number of seconds between each performance sweep, to identify the hot-spot flows.

[0054] A congestion indicator value can be calculated for a remote switch port of an end node based on a formula, $\Delta xmitWait / Interval$. The congestion indicator value defines the normalized port congestion as the number of XmitWaits per second. If the congestion indicator value exceeds a threshold value, it indicates that the endnode is a hot-spot.

5 **[0055]** An oversubscribed end node with a high congestion indicator value is either a contributor to the congestion or a victim flow. For example, the contributors at end node 1, 5, 10 and the victim at end node 2 in Figure 4 all can have a high congestion indicator value. On the other hand, an end node that has a high congestion indicator value for its remote switch port indicates that it is an end point hotspot. For example, the remote switch port that is connected to end
10 node 9 in Figure 4 can have a high congestion indicator value.

[0056] The sender port bandwidth can be measured for each port based on a formula, e.g. $\Delta xmitWait * 4 / Interval$. This formula is derived from the XmitData performance counter that represents the number of bytes transmitted between the performance sweeps. In the formula, the XmitData performance counter is multiplied by 4, because the XmitData counter is measured
15 in a unit of 32-bit words.

[0057] The port utilization can be defined as the ratio between the actual bandwidth and the maximum supported link bandwidth.

[0058] In accordance with an embodiment of the invention, the dFtree implementation includes two algorithms: a first algorithm (e.g. Algorithm 1 as shown below) to identify the hot-spot flows and a second algorithm (e.g. Algorithm 2 as shown below) to reassign a hot-spot flow
20 to a virtual lane classified as a slow lane.

Algorithm 1 Detect endpoint hot-spot and its contributors

Ensure: Subnet is up and running and PM is constantly sweeping

```

1: for  $sw_{src} = 0$  to  $sw_{max}$  do
2:   for  $port_{sw} = 0$  to  $port_{max}$  do
3:     if  $remote\_port(port_{sw}) = HCA$  then
4:       if  $congestion_{port} > Threshold$  then
5:         if  $port_{sw} \neq \text{hot-spot}$  then
6:           Mark  $port_{sw}$  as  $hotspot_{port}$ 
7:         end if
8:         Encapsulate  $hotspot_{LID}$  in a repath trap
9:         Encapsulate  $slow\ lane$  as  $SL_{repath\ trap}$ 
10:      for  $hca_{src} = 0$  to  $hca_{max}$  do
11:        if  $congestion_{port} > Threshold$  then

```

```

12:         if  $hca \neq hotspot_{LID}$  contributor then
13:             if  $Utilisation_{port} < 0.5$  then
14:                 Mark  $hca$  as  $hotspot_{LID}$  contributor
15:                 Forward repath trap to HCA
16:             end if
17:         end if
18:     end if
19: end for
20: else if  $congestion_{port} < \text{Threshold}$  then
21:     if  $port_{sw} = \text{hot-spot}$  then
22:         Clear  $port_{sw}$  as  $hotspot_{port}$ 
23:         Encapsulate  $hotspot_{LID}$  in a unpath trap
24:         Encapsulate fast lane as  $SL_{repath trap}$ 
25:         for  $hca_{src} = 0$  to  $hca_{max}$  do
26:             if  $hca$  is  $hotspot_{LID}$  contributor then
27:                 Clear  $hca$  as  $hotspot_{LID}$ 
28:                 Forward unpath trap to HCA
29:             end if
30:         end for
31:     end if
32: end if
33: end if
34: end for
35: end for

```

Algorithm 2 Reconfigure QP to slow/fast lane

Ensure: Host receives repath trap

```

1: for  $QP_i = 0$  to  $QP_{max}$  do
2:     if  $DLID_{QP} = DLID_{repath trap}$  then
3:         Reconfigure  $SL_{QP}$  according to  $SL_{repath trap}$ 
4:     end if
5: end for

```

[0059] Algorithm 1 can be executed after every iteration of the performance sweep. The algorithm checks if the remote switch port of an end node has a congestion indicator value

exceeding the threshold. For example, the threshold value for congestion that is use to determine congestion can be set as 100000 XmtWait ticks per second. The XmtWait counter is calculated on a per port basis, so the threshold value to determine congestion is applicable even if the network size increases.

- 5 **[0060]** If the remote switch port of an end node has a congestion indicator value exceeding the threshold, then the conclusion is that the end node is a hot spot and the remote switch port is marked as a hot spot port. After discovering an endpoint hot-spot, the first algorithm triggers the forwarding of a repath trap to all potential contributors. This repath trap encapsulates the LID of the congested node.
- 10 **[0061]** The detection of the hot-spot flows depends on the interval of the performance sweeps. If a hot-spot appeared just after iteration n , the hot-spot detection and the 'slow lane' assignment can only be performed at iteration $n + 1$, i.e. t seconds later.
- 15 **[0062]** The congestion indicator value and the port utilization ratio can be used to identify a potential contributor. The congestion indicator value exceeding the threshold indicates that an end node can be either a hot-spot contributor or a victim flow, whereas the port utilization ratio can be used to differentiate between a fair share link and a congested link.
- 20 **[0063]** For example, if node A and node B are sending simultaneously toward node C. Even though both node A and B have a congestion indicator that exceeds the threshold, they receive a fair share of the link bandwidth toward node C. Thus, the algorithm marks an end node as a potential contributor for a hot spot and forwards a repath trap if the congestion indicator value is above the threshold and the port utilization ratio is less than 50%.
- 25 **[0064]** In addition, if a new flow is directed to an existing hot spot, the new flow can be moved to the slow lane. On the opposite, if an end node is no longer a hot-spot, all flows that are directed to that end node can be moved back to its virtual lane classified as fast lane.
- 30 **[0065]** When a repath trap is received by a potential contributor, the Algorithm 2 can be executed. The host can retrieve all the active QPs and compare them with the DLID in the repath trap. If a matching DLID is found in one of the QPs, the QP is reconfigured to use a slow lane. Initially, all QPs are initialized using a fast lane.
- 35 **[0066]** Referring back to Figure 3, the dFtree algorithm can avoid HOL blocking after the PM detects that node 5 is the hot-spot when the congested flows are presented. Then, a repath trap that encapsulates node 5 as a hot-spot LID is forwarded to the source node of the contributors and the victim flows. When a sender (hot-spot contributor or a victim flow) receives the repath trap, the sender retrieves all the active QPs and compares the destination LID with the repath trap LID. If a QP has a matching destination LID, the QP can be reconfigured to the slow lane.
- After the reconfiguration, the victim flow regains its throughput because the dFtree algorithm

places the congested flows in a separated VL (the slow lane) that resolves the HOL blocking.

[0067] The present invention may be conveniently implemented using one or more conventional general purpose or specialized digital computer, computing device, machine, or microprocessor, including one or more processors, memory and/or computer readable storage media programmed according to the teachings of the present disclosure. Appropriate software coding can readily be prepared by skilled programmers based on the teachings of the present disclosure, as will be apparent to those skilled in the software art.

[0068] In some embodiments, the present invention includes a computer program product which is a storage medium or computer readable medium (media) having instructions stored thereon/in which can be used to program a computer to perform any of the processes of the present invention. The storage medium can include, but is not limited to, any type of disk including floppy disks, optical discs, DVD, CD-ROMs, microdrive, and magneto-optical disks, ROMs, RAMs, EPROMs, EEPROMs, DRAMs, VRAMs, flash memory devices, magnetic or optical cards, nanosystems (including molecular memory ICs), or any type of media or device suitable for storing instructions and/or data.

[0069] The foregoing description of the present invention has been provided for the purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise forms disclosed. Many modifications and variations will be apparent to the practitioner skilled in the art. The embodiments were chosen and described in order to best explain the principles of the invention and its practical application, thereby enabling others skilled in the art to understand the invention for various embodiments and with various modifications that are suited to the particular use contemplated. It is intended that the scope of the invention be defined by the following claims and their equivalence.

Claims:

1. A method for alleviating traffic congestion in a middleware machine environment operating on one or more microprocessors, comprising:
 - 5 sweeping, via a subnet manager, a subnet in the middleware machine environment to discover changes and maintain the subnet fully connected;
 - retrieving, via a performance manager, performance and error-related information from one or more performance management agents that are associated with one or more components in the subnet; and
 - 10 allowing a host to dynamically reconfigure addressing state information in order to improve network performance.
2. The method according to Claim 1, further comprising:
 - 15 supporting a fat-tree topology in the middleware machine environment.
3. The method according to Claim 2, further comprising:
 - allowing the fat-tree topology in the middleware machine environment to be a simple fat-tree topology.
- 20 4. The method according to Claim 2, further comprising:
 - allowing the fat-tree topology in the middleware machine environment to be an over-subscribed fat-tree topology.
5. The method according to any preceding Claim, further comprising:
 - 25 using a routing algorithm to ensure deadlock freedom for a packet forwarded from a source node to a destination node.
6. The method according to Claim 5, further comprising:
 - 30 allowing the routing algorithm to forward the packet through a least common ancestor node of the source node and the destination node.
7. The method according to Claim 6, further comprising:
 - 35 allowing the routing algorithm to include an upward stage, in which the packet is forwarded from the source node to the least common ancestor node, and a downward stage, in which the packet is forwarded from the least common ancestor node to the destination.

8. The method according to any preceding Claim, further comprising:
configuring one or more physical links in the middleware machine environment to be associated with multiple virtual lanes.
- 5 9. The method according to any preceding Claim, further comprising:
using one or more performance counter as metrics to identify hot-spots and contributors to the hot-spots during network operation.
- 10 10. The method according to any preceding Claim, further comprising:
identifying a hot-spot flow after a performance sweep.
11. The method according to Claim 10, further comprising:
forwarding a repath message to all contributors after discovering a hot-spot.
- 15 12. The method according to Claim 10 or 11, further comprising:
reassigning the hot-spot flow to a virtual lane classified as a slow lane.
13. The method according to Claim 10 or 11, further comprising:
allowing the hot-spot flow in a simple fat-tree topology to be assigned to a virtual lane
20 classified as a slow lane if a victim flow shares an upward stage with the hot-spot flow.
14. The method according to Claim 10 or 11, further comprising:
allowing the hot-spot flow in an over-subscribed fat-tree topology to be assigned to a
virtual lane classified as a slow lane if a victim flow shares at least one of an upward stage and a
25 downward stage with the hot-spot flow.
15. The method according to Claim 10 or 11, further comprising:
directing all flow to an node from a slow lane to a fast lane, after it is determined that a
node is no longer a hot spot.
30
16. The method according to any preceding Claim, further comprising:
allowing a victim flow to avoid head-of-line blocking using a congestion control mechanism.
- 35 17. A computer program comprising program instructions that when executed by one or more computer systems cause the one or more systems to perform the method of any preceding

claim.

18. A computer program product comprising a machine readable storage medium having the computer program of claim 17 stored thereon.

5

19. A system for preventing traffic congestion in a middleware machine environment operating on one or more microprocessors, comprising:

a subnet manager that sweeps a subnet in the middleware machine environment to discover changes and maintain the subnet fully connected;

10 a performance manager that retrieves performance and error-related information from one or more performance management agents that are associated with one or more components in the subnet; and

a host side stack that can dynamically reconfigure addressing state information in order to improve network performance.

15

20. The system according to Claim 19, wherein:

the subnet in the middleware machine environment is in a fat-tree topology.

21. A non-transitory machine readable storage medium having instructions stored thereon that when executed cause a system to perform the steps comprising:

20

sweeping, via a subnet manager, a subnet in the middleware machine environment to discover changes and maintain the subnet fully connected;

retrieving, via a performance manager, performance and error-related information from one or more performance management agents that are associated with one or more components in the subnet; and

25

allowing a host to dynamically reconfigure one or more virtual lanes in order to improve network performances.

22. A program that when executed causes a system to perform the steps of:

30

sweeping, via a subnet manager, a subnet in the middleware machine environment to discover changes and maintain the subnet fully connected;

retrieving, via a performance manager, performance and error-related information from one or more performance management agents that are associated with one or more components in the subnet; and

35

allowing a host to dynamically reconfigure one or more virtual lanes in order to improve network performances.

100

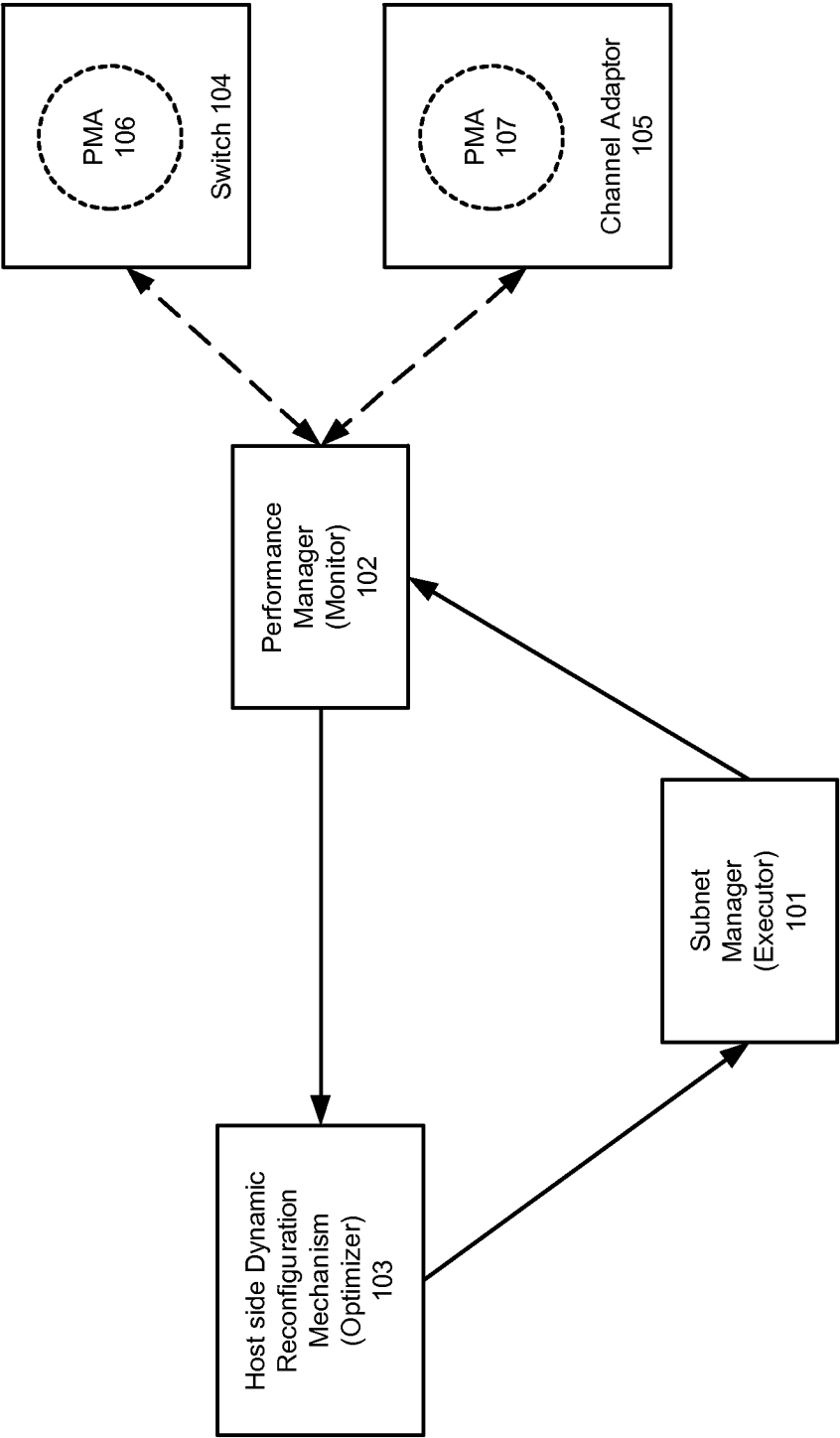


FIGURE 1

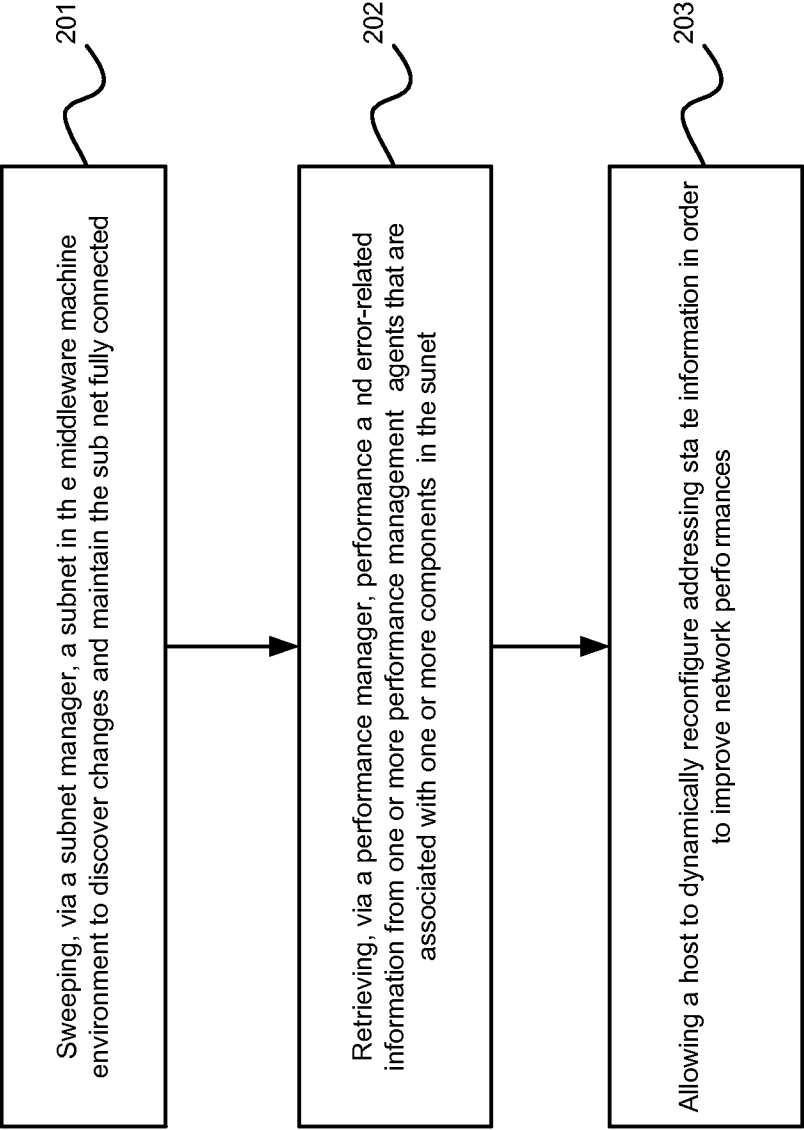


FIGURE 2

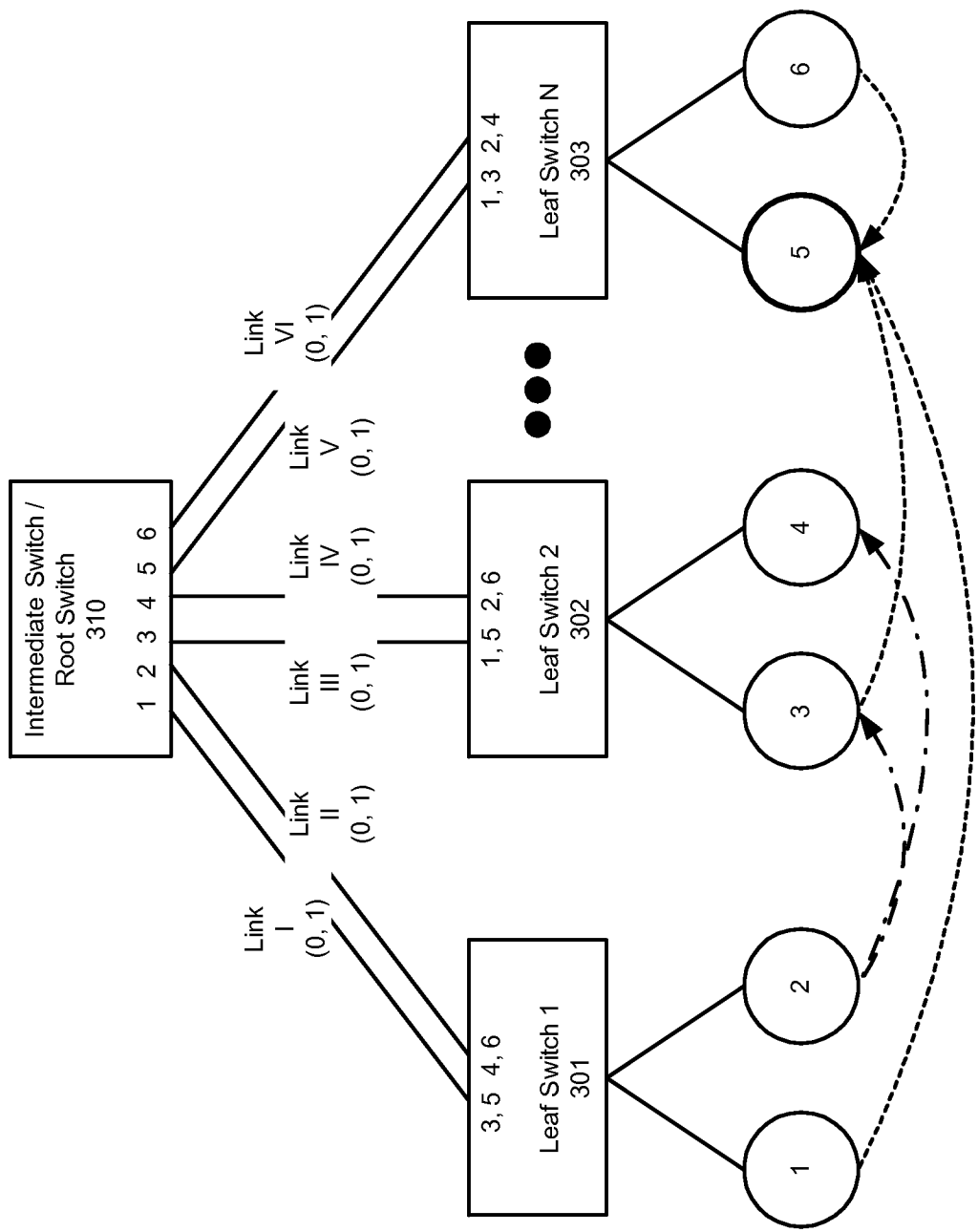


FIGURE 3

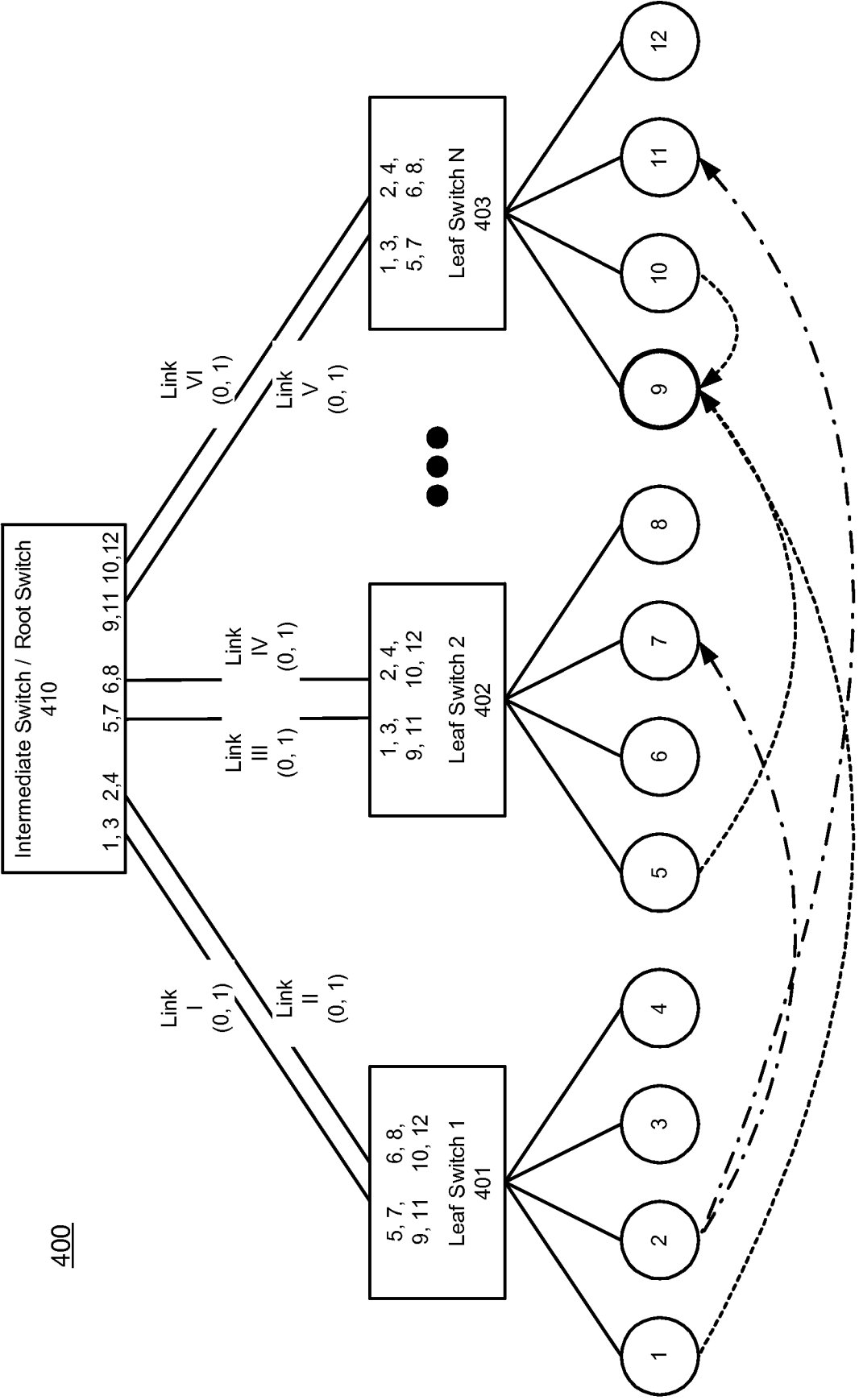


FIGURE 4

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2012/065115

A. CLASSIFICATION OF SUBJECT MATTER INV. H04L12/729 H04L12/707 H04L12/753 ADD.		
According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) H04L		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) EPO-Internal		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Wei Lin Guay, Sven-Arne Reinemo, Olav Lysne, Tor Skeie: "dFtree: a fat-tree routing algorithm using dynamic allocation of virtual lanes to alleviate congestion in infiniband networks", NDM '11 Proceedings of the first international workshop on Network-aware data management, November 14, 2011, Seattle Washington, USA, 14 November 2011 (2011-11-14), pages 1-10, XP002693814, ACM New York, NY, USA ©2011 DOI: 10.1145/2110217.2110219 ISBN: 978-1-4503-1132-8 Retrieved from the Internet: URL:http://delivery.acm.org/10.1145/2120000/2110219/p1-guay.pdf?ip=145.64.134.240&ac c=ACTIVE%20SERVICE&CFID=184667486&CFTOKEN=15819786&__acm__=1363190455_b5697064ff5791 -/--	1-22
☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.		
* Special categories of cited documents : "A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier application or patent but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family		
Date of the actual completion of the international search 14 March 2013		Date of mailing of the international search report 27/03/2013
Name and mailing address of the ISA/ European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016		Authorized officer Plata-Andres, Isabel

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2012/065115

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
	b4a683b3670aec4b59 [retrieved on 2013-03-13] the whole document -----	
X	GUAY W L ET AL: "Host Side Dynamic Reconfiguration with InfiniBand", CLUSTER COMPUTING (CLUSTER), 2010 IEEE INTERNATIONAL CONFERENCE ON, IEEE, PISCATAWAY, NJ, USA, 20 September 2010 (2010-09-20), pages 126-135, XP031774317, ISBN: 978-1-4244-8373-0	1-5, 17-22
Y	the whole document -----	6-16
X	WEI LIN GUAY ET AL: "vFtree - A Fat-Tree Routing Algorithm Using Virtual Lanes to Alleviate Congestion", PARALLEL&DISTRIBUTED PROCESSING SYMPOSIUM (IPDPS), 2011 IEEE INTERNATIONAL, IEEE, 16 May 2011 (2011-05-16), pages 197-208, XP032052376, DOI: 10.1109/IPDPS.2011.28, ISBN: 978-1-61284-372-8	1-5, 17-22
Y	the whole document -----	6-16
A	ABHINAV VISHNU ET AL: "Topology agnostic hot-spot avoidance with InfiniBand", CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE, vol. 21, no. 3, 10 March 2009 (2009-03-10), pages 301-319, XP055056447, ISSN: 1532-0626, DOI: 10.1002/cpe.1359 the whole document -----	1-22